文章编号:1672-3961(2024)04-0051-08 DOI:10.6040/j.issn.1672-3961.0.2023.125

混合 BERT 和宽度学习的低时间复杂度短文本分类

陈晓江[1,2],杨晓奇[2],陈广豪[3],刘伍颖[4,5*]

(1.广东开放大学揭阳分校信息科,广东 揭阳 522095; 2.广东外语外贸大学信息科学与技术学院,广东 广州 510006; 3.广州软件学院软件工程系,广东 广州 510990; 4.鲁东大学山东省语言资源开发与应用重点实验室,山东 烟台 264025; 5.广东外语外贸大学外国语言学及应用语言学研究中心,广东 广州 510420)

摘要:针对短文本分类任务效率低下和精度不高的问题,提出混合基于 Transformer 的双向编码器表示和宽度学习分类器(hybrid bidirectional encoder representations from transformer and broad learning, BERT-BL)的高效率和高精度文本分类模型。对基于 Transformer 的双向编码器表示(bidirectional encoder representation from transformer, BERT)进行微调以更新 BERT 的参数。使用微调好的 BERT 将短文本映射成对应的词向量矩阵,将词向量矩阵输入宽度学习(broad learning, BL)分类器中以完成分类任务。试验结果显示,BERT-BL 模型在 3 个公共数据集上的准确率均达到最优;所需要的时间仅为基线模型支持向量机(support vector machine, SVM)、长短期记忆网络(long short-term memory, LSTM)、最小 p 范数宽度学习(minimum p-norm broad learning, p-BL)和 BERT 的几十分之一,而且训练过程不需要高性能显卡的参与。通过对比分析,BERT-BL 模型不仅在短文本任务中具有良好的性能,而且能节省大量训练时间成本。

关键词:短文本分类; BERT-BL; BERT; 宽度学习; 高精度

中图分类号:TP391 **文献标志码**:A

引用格式:陈晓江,杨晓奇,陈广豪,等. 混合 BERT 和宽度学习的低时间复杂度短文本分类[J].山东大学学报(工学版),2024,54(4):51-58.
CHEN Xiaojiang, YANG Xiaoqi, CHEN Guanghao, et al. Low time complexity short text classification based on fusion of BERT and broad learing [J]. Journal of Shandong University (Engineering Science), 2024, 54(4):51-58.

Low time complexity short text classification based on fusion of BERT and broad learing

CHEN Xiaojiang[1,2], YANG Xiaoqi[2], CHEN Guanghao[3], LIU Wuying[4,5*]

(1. Information Department, Jieyang Campus of Guangdong Open University, Jieyang 522095, Guangdong, China; 2. School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou 510006, Guangdong, China; 3. Department of Software Engineering, Software Engineering Institute of Guangzhou, Guangzhou 510990, Guangdong, China; 4. Shandong Key Laboratory of Language Resources Development and Application, Ludong University, Yantai 264025, Shandong, China; 5. Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, Guangzhou 510420, Guangdong, China)

Abstract: To address the issues of low efficiency and low accuracy in short text classification (STC) tasks, a high-efficiency and high-precision text classification model was proposed that combined transformer based on bidirectional encoder representations and broad learning classifiers (BERT-BL). Through the process of fine-tuning the bidirectional encoder representation from transformer (BERT) based on transformer, the parameters of BERT could be updated to optimize its performance. Utilized fine-tuned BERT to map the short text to its respective word vector matrix, then input it into the BL classifier to classify. The experimental results

收稿日期:2023-06-09

基金项目:教育部新文科研究与改革实践资助项目(2021060049);山东省研究生教育教学改革研究资助项目(SDYJG21185);山东省本科教学改革研究重点资助项目(Z2021323);教育部人文社会科学研究青年基金资助项目(20YJC740062);上海市哲学社会科学"十三五"规划课题资助项目(2019BYY028);教育部人文社会科学研究规划基金资助项目(20YJAZH069);广州市科技计划资助项目(202201010061)

第一作者简介:陈晓江(1995—),男,广东揭阳人,助教,硕士,主要研究方向为自然语言处理。E-mail: 774847467@qq.com

*通信作者简介**:刘伍颖(1980—),男,江西九江人,教授,硕士生导师,博士,主要研究方向为计算语言学和自然语言处理。E-mail: wyliu@ldu.edu.cn

showed that the accuracy of the BERT-BL model reached state-of-art performance on three public datasets respectively. The main finding was that the BERT-BL model took only a few tenths of the time required to baseline models of support vector machine (SVM), long short-term memory (LSTM), minimum p-norm broad learning (p-BL) and BERT, and its training process did not require the participation of a graphics processing unit. Through comparative analysis, the BERT-BL model not only had good performance in STC, but also can save a lot of training time cost.

Keywords: short text classification; BERT-BL; BERT; broad learning; high accuracy

0 引言

文本分类是自然语言处理任务中经典的科学问题,在许多科学和工程中都发挥了重要作用[1]。近年来,文本分类任务在机器学习中引起了广泛关注[2],应用也非常广泛,如数据挖掘、智能交通系统等。基于深度结构的神经网络模型,如卷积神经网络(convolutional neural networks, CNN)和循环神经网络(recurrent neural networks, RNN)广泛用于文本分类任务中,并在性能上突飞猛进。深度神经网络模型通过增加模型隐藏层的数量,使模型结构更加复杂,从而达到更好的性能。尽管深度神经网络模型在许多领域的表现非常出色,但这些模型往往具有大量的超参数和复杂的结构,大多数模型的训练过程耗时过长,对计算机性能提出了更高要求,往往需要高性能设备参与。扁平化的宽度学习(broad learning, BL)神经网络模型不仅解决了深度神经网络训练时间过长的问题,而且在分类性能表现上也不亚于深度神经网络模型。BL 通过在水平方向上增加特征节点组和增强节点组的方式提高模型的性能,由于其简单的模型结构及较少的参数计算,在性能上不仅具有良好的表现,而且具备较低的时间复杂度。BL 结构简单、参数少,已在计算机视觉和图像识别方面取得了显著的效果。深层结构网络模型需在高性能计算机上进行数小时或数天的训练及数百次的迭代,BL 可以在几分钟甚至几秒钟内轻松构建,甚至在一台普通的个人计算机上也能达到良好的性能。BL 最初应用于图像识别任务,在分类性能和时间效率上都具有优越的表现。

文本形式通常可分为长文本和短文本。单字节表示的长文本长度通常超过 255 个字节,而短文本通常指长度相对较短的文本,如聊天信息、新闻话题、观点评论、手机短信等。短文本分类是指按照一定的分类标准对短文进行自动分类和标记,是自然语言处理(natural language processing, NLP)中一个典型而复杂的科学问题[3],面临标记较少、含糊不清、信息不规范等挑战。因此,高精度的短文本分类研究是 NLP 中一个重要组成部分。

文本分类通常分为 2 个步骤:将预处理好的文本转化为计算机可计算的特征向量,也称为文本表示或词嵌入;将特征向量输入分类器中,得到文本分类结果。文本表示是决定文本分类结果表现的关键部分。传统的文本表示方法(如词袋技术)忽略了文本的上下文关系,将词与词之间视为相互独立的存在,导致语义关联较弱和“维度灾难”问题。近年来,神经网络广泛用于词嵌入以获得文本特征,如神经网络语言模型(neural probabilistic language model, NNLM)[4]、Word2vec[5-6]、全局词频统计的词表征(global vectors for word representation, GloVe)[7]。这些方法不仅很好地解决了“维度灾难”问题,还可以学习文本的上下文关系,衡量词与词之间的相似度。上述方法尽管在许多自然语言理解任务中取得了大量成果,但存在另一个潜在问题,即不能区分多义词。基于 Transformer 的双向编码器表示(bidirectional encoder representation from transformer, BERT)在纯文本预训练的基础上,创新性地加入了掩码语言模型(masked language model, MLM)和下句预测任务(next sentence prediction, NSP),在一定程度上能较好地解决多义词的问题[8]。

本研究证明了宽度学习在自然语言处理任务上的有效性,并在 3 个公开短文分类数据集上取得了优异的表现;在 3 个公开数据集上对 BERT 进行微调,充分挖掘其将文本转换成特征向量的潜力。

1 相关工作

在自然语言处理领域,深度学习方法的大部分工作涉及通过神经网络模型计算文本向量表示。

文本向量表示在短文本分类中起着至关重要的作用。NLP 研究人员广泛研究了各种预训练模型,如 NNLM、word2vec 和 GloVe,这些预训练模型可用于生成连续密集的文本向量表示。NNLM 的方法是用前 $n-1$ 个词预测第 n 个词的概率。Word2vec 主要分为 2 种方法:连续词袋(continuous bag of words,CBOW)和 Skip-Gram。在 CBOW 方法中,一个词的向量由该词上下文相关的词向量预测;在 Skip-Gram 方法中,一个词的上下文词向量由该词的向量预测。GloVe 是一个基于全局词频统计的预训练模型,可以将一个单词表示为一个向量,该向量捕捉单词之间的共现关系。这些预训练模型利用神经网络将稀疏的独热编码转换为稠密的词向量,避免“维度灾难”问题,且可以更好地考虑上下文语义信息。这些预训练模型的不足同样明显,它们生成的词向量表示是静态的,即词和对应词向量之间只有一对一的对应关系,导致不能解决多义词表征问题,例如在同一语料库中,“苹果”一词对应的词向量仅有一个,难以对其不同的词义加以区分。为了解决上述问题,文献[8]提出 BERT,在文本特征表示和多义词方面显示出更好的效果。

对于短文本分类任务,许多传统的机器学习方法陆续提出[9],如支持向量机(support vector machine, SVM)[10]、朴素贝叶斯(naive bayes, NB)[11]和 K 最邻近分类(K-nearest neighbor, KNN)[12]算法。SVM 是一个稀疏且健壮的线性分类器,在特征空间上定义了最大区间,是一个二元分类模型;NB 是一种利用概率和统计知识通过结合先验和后验概率对样本数据集进行分类的方法;KNN 算法的核心思想是如果一个样本在特征空间的 K 个最近邻居中的大多数属于一个特定的类别,那么这个样本本身也属于这个类别。由于基于机器学习的方法需要大量的特征工程,近年来,基于深度学习的方法在短文本分类中得到广泛应用[13-15],并展现出良好的效果,如 CNN[16]、胶囊网络(capsule network, CapsNet)[17]和图卷积神经网络(graph neural network, GNN)[18]。CNN 是一种具有卷积计算和深度结构的前馈神经网络,是深度学习的代表算法之一,包含卷积层、池化层和全连接层,可以有效减少特征维度;CapsNet 在 CNN 的基础上加以改进,优化了卷积操作[19];GNN 聚集了每个节点和其周围节点的信息。这些基于深度学习的方法可以达到很好的性能,但需要进行大量的参数计算,耗费大量的训练时间和计算资源。深度学习通常采取增加层数等方法,使模型结构更加复杂,进而具有更好的泛化能力,但需要计算的参数数量和训练时间也会大大增加。为应对传统机器学习和深度学习带来的挑战,文献[20]提出了 BL,通过水平方向上增加节点以提高模型的泛化能力,不仅花费时间少,而且尽可能地保存数据的特征。特征节点和增强节点都是 BL 的组成部分。首先,输入的特征被线性转化为映射特征,生成特征节点;其次,映射得到的特征节点被非线性地转化为具有随机生成权重的增强特征,以生成增强节点。作为一个平面神经网络,BL 具有结构简单、参数少的优点,已广泛用于各种分类任务,如图像分类[21-22]和视觉识别[23]。由于短文本特征比图像特征更密集,BL 在短文本分类任务上的研究相对较少。一些研究者改进了 BL,提出门控宽度学习(gated broad learning system, G-BLS)和递归宽度学习(recurrent broad learning system, R-BLS)模型[24],其原理是在 BL 的节点中加入门控机制和循环机制,使其具有上下文记忆能力。这 2 个模型使用 Word2vec 获得文本的词向量表示,并应用 G-BLS 和 R-BLS 预测文本标签。本研究提出 BERT-BL 模型,通过微调后的 BERT 预训练模型初始化短文本的词向量表示,使 BERT 对 BL 分类起到一定的增强作用,通过 BL 预测分类标签。试验结果表明,BERT-BL 模型可以在更短的时间内达到更先进的性能。

2　BERT-BL 模型

BERT-BL 模型在文本特征提取和快速分类方面具备较优越的性能。数据集中所有短文本都需要进行预处理,具体为截取或填充。本研究需要将文本规范化为合适且相同的长度,用微调后的 BERT 获得文本向量,提取 BERT 最后一个隐藏层特征向量作为文本的句向量矩阵,将这一句向量矩阵输入 BL 分类器以预测文本标签。本章将从 BERT-BL 模型整体入手,分别介绍 BERT 预训练模型和 BL 分类器。BERT-BL 模型如图 1 所示。

2.1　BERT 增强层

在 BERT-BL 模型中,BERT 预训练模型用于嵌入层以计算文本表征。BERT 结构如图 2 所示。

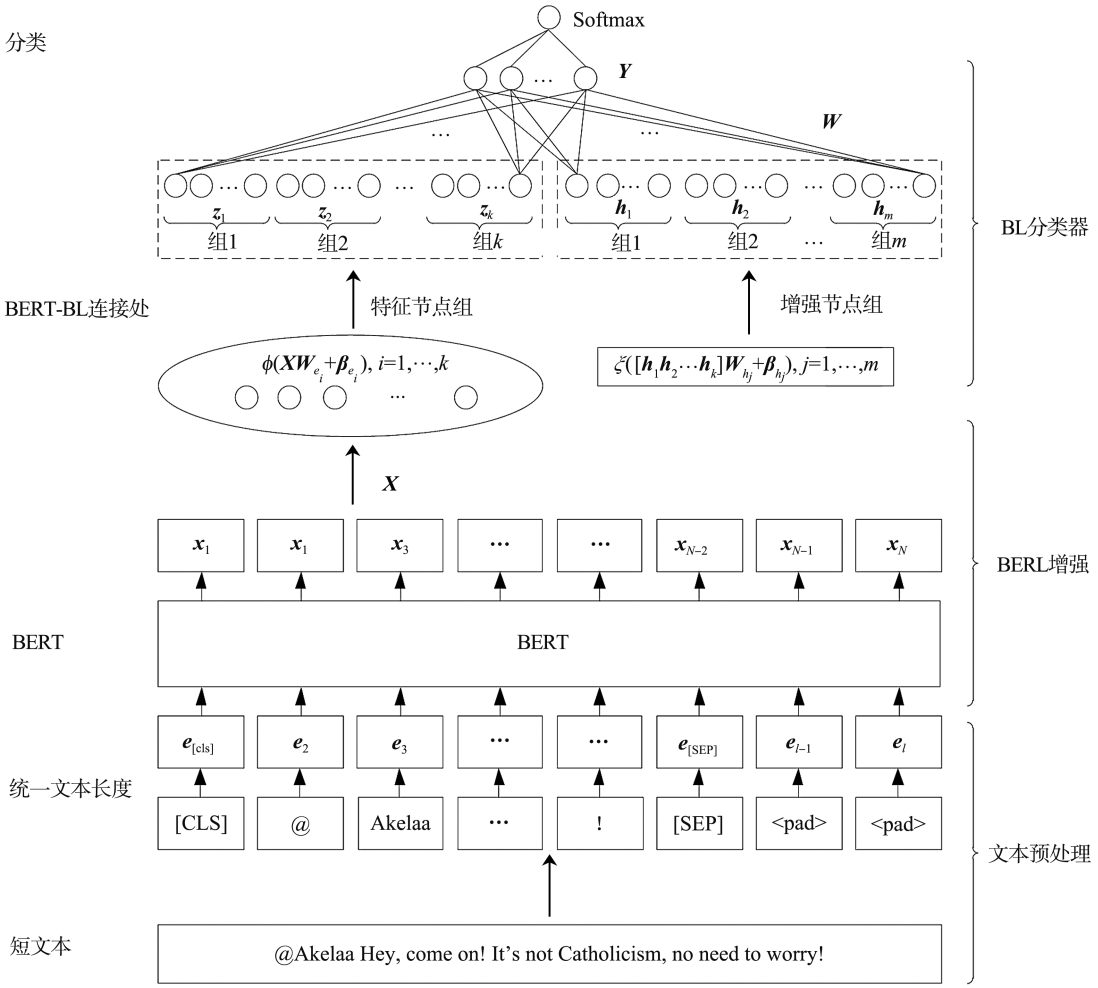

图 1 BERT-BL 模型示意图
Fig.1 BERT-BL model

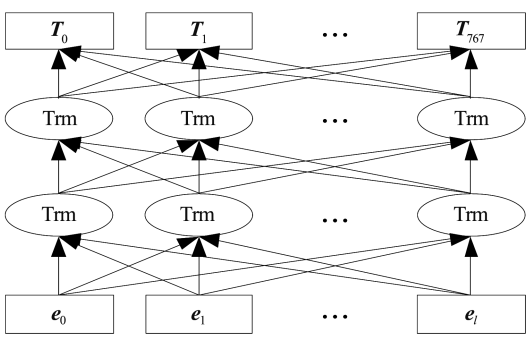

图 2 BERT 结构示意图
Fig.2 BERT model

图 2 中 $\boldsymbol{e}_l$ 为文本的词嵌入矩阵,l 为文本经过截取或填充后的统一长度,Trm 为 Transformer 块编码器,$\boldsymbol{T}_i(i=0,1,\cdots,767)$ 为 BERT 最后一个隐藏层向量。BERT 包含 12 个 Transformer 块编码器、12 个自注意力机制头,最后一个隐藏层维度为 768。Transformer 最重要的部分是用自注意机制取代 RNN,使其在捕捉语义关联上更加高效。Transformer 使用自注意机制获得单词之间的关系,反映每个单词在句子中的权重。对于同一个词,不同的上下文会让这个词融合不同的语义信息,使同一个词在不同的上下文中有不同的词向量,表征不同的语义,从而区分一词多义。自注意机制的输出向量

$$\boldsymbol{A}=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}, \tag{1}$$

式中:d_k 为输入向量的维度;$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 分别为查询向量、关键向量和内容向量,其概念取自信息检索系统。类比一个简单的搜索例子,当人们在检索平台上搜索一个产品时,在搜索引擎上输入内容 $\boldsymbol{Q}$,搜索引擎根据 $\boldsymbol{Q}$ 与关键词 $\boldsymbol{K}$ 进行匹配,根据 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 的相似度得到匹配的内容 $\boldsymbol{V}$,根据归一化指数函数得到注意力层的输出向量 $\boldsymbol{A}$。

在通用领域中训练 BERT,虽然可以赋予 BERT 很强的泛化能力,但也可能导致 BERT 在特定目标领域表现欠佳。为了解决这一问题,本研究用目标域数据对 BERT 做进一步预训练,即微调 BERT,将 BERT 输出的最后一层隐藏层向量作为句子向量与 BL 分类器相连接,通过计算相关权重预测短文本的标签。

2.2　BL 分类器

BL 分别器是一种不依赖深度结构的神经网络模型,具有很高的运行效率和简单的结构,其结构如图 3 所示。

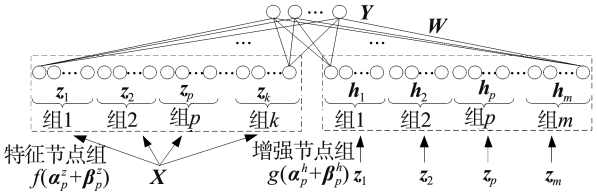

图 3　BL 分类器

Fig.3　BL classifier

BERT 输出的向量矩阵 $\boldsymbol{X}=[\boldsymbol{x}_1^{\mathrm{T}}\quad \boldsymbol{x}_2^{\mathrm{T}}\quad \cdots\quad \boldsymbol{x}_N^{\mathrm{T}}]^{\mathrm{T}}\in\mathbf{R}^{N\times M}$通过线性变换转化为 BL 的输入,其中 N 为数据集样本数量,M 为向量 $\boldsymbol{x}_i^{\mathrm{T}}(i=1,2,\cdots,N)=[\boldsymbol{T}_0\quad \boldsymbol{T}_1\quad \cdots\quad \boldsymbol{T}_M]$的维度。$\boldsymbol{Y}=[\boldsymbol{y}_1^{\mathrm{T}}\quad \boldsymbol{y}_2^{\mathrm{T}}\quad \cdots\quad \boldsymbol{y}_N^{\mathrm{T}}]^{\mathrm{T}}\in\mathbf{R}^{N\times C}$为标签矩阵,其中 C 为标签的类别数,也是标签向量 $\boldsymbol{y}_i^{\mathrm{T}}(i=1,2,\cdots,N)$ 的维度。$\boldsymbol{Z}^k=[\boldsymbol{z}_1\quad \boldsymbol{z}_2\quad \cdots\quad \boldsymbol{z}_k]\in\mathbf{R}^{N\times(k\times N_M)}$为特征节点矩阵,通过 BERT 输出的向量矩阵线性转化而来,可以通过试验设置合适的节点组数,尽可能保留数据原本的特征,从而更好地拟合模型,其中 k 为特征节点的组数,每个特征节点组 $\boldsymbol{z}_p(p=1,2,\cdots,k)$ 包含 N_M 个隐藏节点,$\boldsymbol{z}_p=f(\boldsymbol{X}\boldsymbol{\alpha}_p^z,\boldsymbol{\beta}_p^z)$,其中 f 为激活函数(如 Relu),$\boldsymbol{\alpha}_p^z\in\mathbf{R}^{M\times N_M}$和$\boldsymbol{\beta}_p^z\in\mathbf{R}^{N\times N_M}$分别为随机生成的权重和偏置,用于映射得出特征节点组 z_p。$\boldsymbol{H}^m=[\boldsymbol{h}_1\quad \boldsymbol{h}_2\quad \cdots\quad \boldsymbol{h}_m]\in\mathbf{R}^{N\times(m\times N_E)}$为增强节点矩阵,通过特征节点矩阵非线性变换转化而来,目的是增强模型非线性复杂度,其中 m 为增强节点的组数,每个节点组 $\boldsymbol{h}_p(p=1,2,\cdots,m)$ 都有 N_E 个隐藏节点,$\boldsymbol{h}_p=g(\boldsymbol{h}_p\boldsymbol{\alpha}_p^h,\boldsymbol{\beta}_p^h)$,其中 g 为激活函数,可以与 f 相同,$\boldsymbol{\alpha}_p^h\in\mathbf{R}^{N_M\times N_E}$和$\boldsymbol{\beta}_p^h\in\mathbf{R}^{N\times N_E}$分别为随机生成的权重和偏置,用于映射得出 $\boldsymbol{h}_p$。综上,$\boldsymbol{z}_p$ 和 $\boldsymbol{h}_p$ 的维度分别为 $N\times N_M$ 和 $N\times N_E$。训练为文本的真实标签

$$\boldsymbol{Y}=[\boldsymbol{z}_1\quad \cdots\quad \boldsymbol{z}_k\mid \boldsymbol{h}_1\quad \cdots\quad \boldsymbol{h}_m]\boldsymbol{W},\tag{2}$$

式中 $\boldsymbol{W}$ 为输出层权重。BL 的核心目标是寻找一个 $\boldsymbol{W}$ 使预测标签 $\hat{\boldsymbol{Y}}$ 和真实标签 $\boldsymbol{Y}$ 之间的差异尽可能小。因此,上述问题可以转化为:

$$\arg\min_{\boldsymbol{W}}\left(\|\boldsymbol{Y}-\hat{\boldsymbol{Y}}\|_2^2+\frac{\lambda}{2}\|\boldsymbol{W}\|_2^2\right),\tag{3}$$

$$\boldsymbol{W}=\lim_{x\to 0}(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}+\lambda\boldsymbol{I})^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{Y},\tag{4}$$

式中:λ 为正则化参数;$\boldsymbol{U}$ 为特征节点组和增强节点组的拼接,$\boldsymbol{U}=[\boldsymbol{Z}^k\mid\boldsymbol{H}^m]\in\mathbf{R}^{N\times(k\times N_M+m\times N_E)}$,在实际情况中 $\boldsymbol{U}$ 往往是没有逆或求不出逆的,但可通过求伪逆的方式求解 $\boldsymbol{W}$;$\boldsymbol{I}$ 为单位矩阵。

2.3　BERT-BL 算法

BERT-BL 构建模型和训练的过程如算法 1 所示。

算法 1　BERT-BL 的训练步骤

输入: 短文本数据集 $T_{\mathrm{ext}}=\{t_1\ t_2\cdots t_N\}$、标签 $\boldsymbol{Y}$。

输出: BL 输出层权重 $\boldsymbol{W}$。

① **for** $i=1\to N$ **do**

②　$\boldsymbol{E}_i\leftarrow$tokenizer$(t_i)$;

③ **end for**

④ 可得短文本的词嵌入矩阵 $\boldsymbol{E}=[\boldsymbol{E}_1\quad \boldsymbol{E}_2\quad \cdots\quad \boldsymbol{E}_N]$;

⑤ **for** $j=1\to N$ do

⑥　$\boldsymbol{x}_i\leftarrow$BERT$(\boldsymbol{E}_i)$;

⑦ **end for**

⑧ 可得 BERT 输出为 $\boldsymbol{X}=[\boldsymbol{x}_1^{\mathrm{T}}\quad \boldsymbol{x}_2^{\mathrm{T}}\quad \cdots\quad \boldsymbol{x}_N^{\mathrm{T}}]^{\mathrm{T}}\in\mathbf{R}^{N\times M}$;

⑨ **for** $i=1\to k$ **do**

⑩　随机初始化$\boldsymbol{\alpha}_p^z\in\mathbf{R}^{M\times N_M}$和$\boldsymbol{\beta}_p^z\in\mathbf{R}^{N\times N_M}$;

⑪　$\boldsymbol{z}_i\leftarrow f(\boldsymbol{X}\boldsymbol{\alpha}_i^z+\boldsymbol{\beta}_i^z)$;

⑫　**end for**

⑬ 可得 k 组特征节点 $\boldsymbol{Z}^k=[\boldsymbol{z}_1\quad \boldsymbol{z}_2\quad \cdots\quad \boldsymbol{z}_k]\in\mathbf{R}^{N\times(k\times N_M)}$;

⑭ **for** $i=1\to m$ **do**

⑮　随机初始化$\boldsymbol{\alpha}_p^h\in\mathbf{R}^{N_M\times N_E}$和$\boldsymbol{\beta}_p^h\in\mathbf{R}^{N\times N_E}$;

⑯　$\boldsymbol{h}_i\leftarrow g(\boldsymbol{Z}\boldsymbol{\alpha}_p^h+\boldsymbol{\beta}_p^h)$;

⑰　**end for**

⑱ 可得 m 组增强节点 $\boldsymbol{H}^m=[\boldsymbol{h}_1\quad \boldsymbol{h}_1\quad \cdots\quad \boldsymbol{h}_m]$;

⑲ $\boldsymbol{U}\leftarrow[\boldsymbol{Z}^k\mid\boldsymbol{H}^m]$;

⑳ 根据式(4)算出输出层的权重 $\boldsymbol{W}$。

算法中 t_i 为短文本数据集 T_{ext} 中的第 i 句短文本,tokenizer 为 BERT 的分词器,$\boldsymbol{E}_i=[\boldsymbol{e}_0\ \boldsymbol{e}_1\cdots\boldsymbol{e}_l]$为文本通过分词器处理后得到的词嵌入矩阵。

3　试验

为证实所提模型的有效性,本研究在 3 个公开的英文公共短文本数据集上进行试验。此外,将 BERT-BL 与流行的基线模型进行比较,以验证本研究所提模型的效率和性能。在 BERT 选择上,本研究采用 Bert-base-uncased 开源模型,该模型不区分大小写。

3.1　评价指标

本研究采用 2 个较为常用的评价指标,分别是准确率和训练时间,其中准确率

$$A_{\mathrm{ccuracy}}=\frac{T_{\mathrm{P}}+T_{\mathrm{N}}}{T_{\mathrm{P}}+T_{\mathrm{N}}+F_{\mathrm{P}}+F_{\mathrm{N}}},\tag{5}$$

式中,T_P 为模型预测正确的正样本个数,T_N 为模型预测正确的负样本个数,F_P 为模型预测错误的正样本个数,F_N 为模型预测错误的负样本个数。

3.2 数据集

本研究在 3 个公开英文短文本数据集上开展试验,分别为 Apple-Twitter-Sentiment(ATS)数据集、AirlineSentiment(AS)数据集和 Tweet-global-warming(Tgw)数据集。这 3 个数据集的主题类别主要为话题和情感分析,每个数据集均随机取 80% 的文本作为训练集,20% 的文本作为测试集。数据集的属性如表 1 所示。

表 1 数据集属性

Table 1 Properties of datasets

数据集	样本数/个	文本最大长度	类别数/个
ATS	3 886	37	4
AS	14 606	35	3
Tgw	6 090	41	3

3.3 基线模型

本研究以对比的方式评估 BERT-BL 模型的性能,使用以下模型作为基线模型。

(1) R-BLS[24]。为解决 BL 模型在学习序列信息和单词重要性方面的局限性,提出 R-BLS 模型。该模型和 RNN 相似,同时利用了序列信息和单词的重要性。

(2) G-BLS[24]。能同时学习序列信息和词语权重,设计了一个外部遗忘门以控制模型学习到的序列信息。

(3) SVM[10]。是一个线性分类器。对于线性不可分的样本,SVM 能在更高维度上实现线性划分。同时,核函数的选取对 SVM 也很重要,经过试验,本研究选取“linear”作为 SVM 的核函数。在分类策略上采取“一对多”的策略。

(4) NB[11]。是一种基于贝叶斯定理和特征条件独立性假设的分类方法。将拉普拉斯平滑参数设置为 0.5。

(5) BERT-linear[8]。是一个强大的文本表示模型,结合了 BERT 和全连接层,以全连接层作为分类器。批次大小设定为 64,学习率设定为5×10^{-5},全连接层的隐藏单元数量为 128,训练周期为 5。

(6) 长短期记忆(long short-term memory, LSTM)网络[25]。是一个经典的深度学习神经网络,可以学习数据的时序序列特征。本研究将 BERT 的输出连接到 LSTM 网络,将其特征向量压缩到 128 维,并最终通过全连接层网络进行分类。批次大小设置为 32,网络层数为 1,隐藏单元数为 256,学习率为5×10^{-5},训练周期为 5。

(7) 最小 p 范数宽度学习(minimun p-norm broad learning, p-BL)[26]:一种使用估计误差 p 次方对 BL 模型最小二乘法进行改进的分类器。每组特征节点数和增强加点数均为 100 个,分别设置了 15 组特征节点和 10 组增强节点,正则化系数 λ 为 20,激活函数选用 tanh 函数。

3.4 试验参数

深度结构的神经网络通常通过增加隐藏层从数据中提取更深层次的特征。然而,需要计算的参数量会大大增加,也会忽略一些分类不明显的特征。虽然有些特征并不明显,但对分类的表现有一定影响。在 BL 中,用特征节点将 BERT 生成的文本向量转化为映射的特征。通过这种方式,BL 可以尽可能地提取和保留文本特征。此外,可以将特征节点进一步转换为增强节点,从而提取更深层次的特征。不同数量的特征节点和增强节点对不同的数据集有不同的分类效果。本研究对每个数据集进行 100 次试验,以了解特征节点和增强节点对分类性能的影响。在 3 个数据集上,不同数量的 2 种节点对应的试验结果如图 4~6 所示。

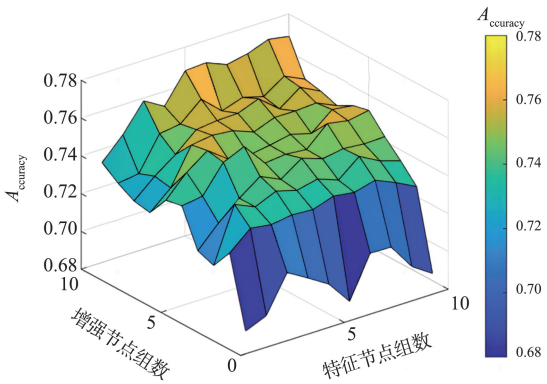

图 4 特征节点和增强节点在 ATS 数据集上对 BL 分类性能的影响

Fig.4 The influence of feature and enhancement nodes on BL classification performance on ATS dataset

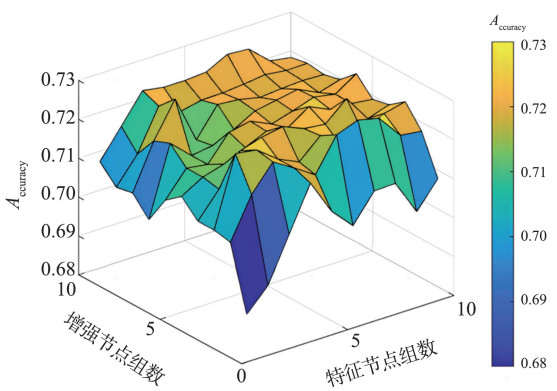

图 5 特征节点和增强节点在 Tgw 数据集上对 BL 分类性能的影响

Fig.5 The influence of feature and enhancement nodes on BL classification performance on Tgw dataset

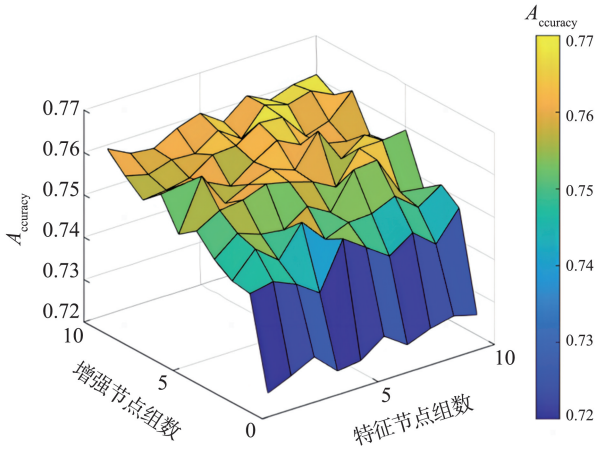

图 6　特征节点和增强节点在 AS 数据集上对 BL 分类性能的影响

Fig.6　The influence of feature and enhancement nodes on BL classification performance on AS dataset

由图 4~6 可知,每组有 100 个节点,特征节点和增强节点的数量对分类准确率有很大影响;随着特征节点和增强节点数量的增加,准确率趋于上升。由图 5 可知:这 2 类节点的数量并非越多越好,当 2 种节点达到一定数量时,可能达到一个相对较好的分类结果,而当 2 种节点的数量过多时,准确率难以提高甚至下降。

BL 的试验参数见表 2,可以不断调整特征节点和增强节点的数量优化模型。同时,正则化参数对试验结果也有很大影响。此外,在使用 BERT 提取文本特征前,本研究对 BERT 进行了微调,微调参数为:批处理量为 32,最大序列长度为 100,学习率为 5×10^{-5},训练步骤为 3 000,预热步骤为 5 000。

表 2　BL 参数

Table 2　Parameters of BL

数据集	特征节点组数	增强节点组数	每组节点个数	正则化参数
ATS	8	5	100	20
AS	8	5	100	1
Tgw	5	3	100	50

3.5　试验结果和分析

本研究使用传统的机器学习和深度学习方法进行对比试验,同时与 *p*-BL、R-BLS 和 G-BLS 的试验结果相比较。试验结果如表 3 所示,BERT-BL 在 3 个数据集上的准确率均达到最高,且训练时间比其他基于深度结构神经网络模型要短得多。R-BLS 和 G-BLS 旨在通过学习序列信息和单词的重要性增强上下文特征,但 R-BLS 和 G-BLS 利用 Word2vec 作为预训练模型获得文本表示,无法解决多义词问题,导致精度不如 BERT-BL。在 NB 方法中,每个特征的计算是独立、并行的,所以 NB 的计算速度比其他方法快。SVM 是一种传统的机器学习模型,最初用于二元分类,难以适应多元分类。虽然 *p*-BL 是对 BL 中的最小二乘法进行改进,但从试验效果上看,*p*-BL 与 BERT-BL 在分类性能上相差无几,但 *p*-BL 的训练时间却比 BERT-BL 高很多。试验结果表明,本研究提出的模型在分类性能和时间复杂度上都具有较优异的表现。

表 3　BERT-BL 试验对比结果

Table 3　Comparison results of BERT-BL experiment

数据集	模型	准确率/%	训练时间/s
ATS	R-BLS	67.34	0.11
	G-BLS	67.76	0.15
	NB	57.94	**0.01**
	SVM	61.36	1.61
	LSTM	70.32	143.00
	BERT	66.97	97.00
	p-BL	76.60	37.16
	BERT-BL	**76.64**	1.04
AS	R-BLS	75.05	6.45
	G-BLS	75.26	9.50
	NB	50.70	**0.02**
	SVM	73.26	26.54
	LSTM	75.80	270.00
	BERT	72.82	156.00
	p-BL	76.00	340.06
	BERT-BL	**76.10**	2.48
Tgw	R-BLS	60.14	6.45
	G-BLS	60.59	9.50
	NB	66.83	**0.02**
	SVM	63.25	26.54
	LSTM	67.89	270.00
	BERT	70.89	156.00
	p-BL	72.43	74.87
	BERT-BL	**72.44**	2.48

4　结论

本研究提出一个低时间复杂度的 BERT-BL 模型解决短文本分类问题。在 3 个公共英文短文本数据集上进行试验,并将 BERT-BL 模型的试验结果与传统的机器学习和深度学习方法试验结果进行比较。试验结果表明,BERT-BL 模型可以在较短时间内达到优越的性能。

随着 BERT-BL 模型的引入,本研究希望

BERT-BL模型作为一个高准确率和高效率的短文分类模型，在现实场景中更具吸引力。未来将专注于解决更多语种的短文本问题，并进一步优化BL分类器，以提高其对短文本分类等NLP任务的性能。

参考文献：

[1] LUO X. Efficient English text classification using selected machine learning techniques[J]. Alexandria Engineering Journal, 2021, 60(3): 3401-3409.

[2] Al-SALEMI B, AYOB M, NOAH S. Feature ranking for enhancing boosting-based multi-label text categorization [J]. Expert Systems with Applications, 2018, 113: 531-543.

[3] SORA O, JUNYA T, TOMOYUKI K, et al. Text classification with negative supervision[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. [S. l.]: Association for Computational Linguistics, 2020: 351-357.

[4] BENGIO Y, SCHWENK H, SENÉCAL J, et al. Neural probabilistic language models[J]. The Journal of Machine Learning Research, 2003, 3(6): 1137-1155.

[5] LE Q, MIKOLOV T. Distributed representations of sentences and documents[C]//Proceedings of the 31st International Conference on International Conference on Machine Learning. Beijing, China: PMLR, 2014: 1188-1196.

[6] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[C]// Proceedings of the 1st International Conference on Learning Representations. Scottsdale, USA: ICLR, 2013: 1-12.

[7] PENNINGTON J, SOCHER R, CHRISTOPHER D M. Glove: global vectors for word representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, 2014: 1532-1543.

[8] DEVLIN J, CHANG M, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, USA: Association for Computational Linguistics, 2019: 4171-4186.

[9] PANG B, LEE L, VAITHYANATHAN S. Thumbs up? Sentiment classification using machine learning techniques [C]//Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP). Philadelphia, USA: Association for Computational Linguistics, 2002: 79-86.

[10] CORTES C, VAPNIK V. Support-vector networks[J]. Machine Learning, 1995, 20(1): 273-297.

[11] BLACK T C, THOMPSON W J. Bayesian data analysis [J]. Computing in Science Engineering, 2001, 3(4): 86-91.

[12] ABEYWICKRAMA T, CHEEMA M A, TANIAR D. *K*-nearest neighbors on road networks: a journey in experimentation and in memory implementation[J]. Proceedings of the VLDB Endowment, 2016, 6(9): 492-503.

[13] SERGIO G C, LEE M. Stacked DeBERT: all attention in incomplete data for text classification[J]. Neural Networks, 2021, 136: 87-96.

[14] MENG Y, ZHANG Y, HUANG J, et al. Text classification using label names only: a language model self-training approach[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). [S. l.]: Association for Computational Linguistics, 2020: 9006-9017.

[15] ZHOU P, QI Z, ZHENG S, et al. Textclassification improved by integrating bidirectional LSTM with two-dimensional max pooling[C]//Proceedings of the Conference on Computational Linguistics. Osaka, Japan: Association for Computational Linguistics, 2016: 3485-3495.

[16] KIM Y. Convolutionalneural networks for sentence classification[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, 2014: 1746-1751.

[17] ZHAO W, YE J, YANG M. Investigating capsule networks with dynamic routing for text classification [C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018: 3110-3119.

[18] CHU Y, LIN H, YANG L, et al. Hyperspectral image classification based on discriminative locality preserving broad learning system[J]. Knowledge-Based Systems, 2020, 206: 106319.

[19] 戴宏，盛立杰，苗启广. 基于胶囊网络的对抗判别域适应算法[J].计算机研究与发展，2021，58(9)：1997-2012.
DAI Hong, SHENG Lijie, MIAO Qiguang. Adversarial discriminative domain adaptation algorithm with CapsNet [J]. Journal of Computer Research and Development, 2021, 58(9): 1997-2012.

（下转第66页）